A Multi-Stream Approach for Seizure Classification with Knowledge Distillation

Jen-Cheng Hou, Aileen Mcgonigal, Fabrice Bartolomei, Monique Thonnat

A Multi-Stream Approach for Seizure Classification with Knowledge Distillation

Jen-Cheng Hou
University of Côte d'Azur, INRIA
Sophia-Antipolis, France
jen-cheng.hou@inria.fr

Aileen McGonigal
Aix-Marseille University
Marseille, France
aileen.MCGonigal@univ-amu.fr

Fabrice Bartolomei
Aix-Marseille University
Marseille, France
Fabrice.BARTOLOMEI@univ-amu.fr

Monique Thonnat
University of Côte d'Azur, INRIA
Sophia-Antipolis, France
Monique.Thonnat@inria.fr

Abstract

In this work, we propose a multi-stream approach with knowledge distillation to classify epileptic seizures and psychogenic non-epileptic seizures. The proposed framework utilizes multi-stream information from keypoints and appearance from both body and face. We take the detected keypoints through time as spatio-temporal graph and train it with an adaptive graph convolutional networks to model the spatio-temporal dynamics throughout the seizure event. Besides, we regularize the keypoint features with complementary information from the appearance stream by imposing a knowledge distillation mechanism. We demonstrate the effectiveness of our approach by conducting experiments on real-world seizure videos. The experiments are conducted by both seizure-wise cross validation and leave-one-subject-out validation, and with the proposed model, the performances of the F1-score/accuracy are 0.89/0.87 for seizure-wise cross validation, and 0.75/0.72 for leave-one-subject-out validation.

1. Introduction

Epilepsy is a neurological disorder, resulting from abnormal electrical discharging in the brain. About 1% of the population worldwide suffer from this disabling condition [34]. The cardinal feature of epilepsy is the tendency to present epileptic seizures (ES), which are often associated with motor activity changes, including repeated or rhythmic movements. Nevertheless, not all seizures are epileptic seizures (ES). Indeed, psychogenic non-epileptic seizures (PNES) are not caused by epileptic neuronal activity in the brain, and are considered to be mainly caused by psychological factors [11, 21]. The clinical management of ES and PNES is different and as such, accurate diagnosis is crucial to avoid therapeutic errors. To diagnose the type of seizure, one important information comes from semiology [31], i.e., the clinical signs that occur during the seizure, independently from auxiliary information such as EEG or neuroimaging. The gold standard diagnostic method is to record habitual events on video-EEG, with simple visual analysis by an expert in epileptology. Nevertheless, distinguishing between ES and PNES may be challenging, with low accuracy rates for less experienced clinicians, especially when seizures of either type involve complex hyperkinetic motor behavior [31]. There have been many works trying to deal with seizure classification problems with machine learning based on either EEG signals [28, 18, 30] or visually observed semiology [16, 3, 26]. However, to our knowledge, none so far have specifically focused on distinguishing ES from PNES.

In this work, we take advantage of recent deep learning frameworks in computer vision for directly analyzing patients' semiology, focusing particularly on the body pose and face regions. Several related works have been proposed recently [2, 4, 17]. In [2], the authors use semiological signs from face, body, and hands to classify epilepsy with convolution neural networks (CNNs) and recurrent neural networks (RNNs). The work in [4] also utilized similar strategy with pre-trained CNN features combined with RNNs for analyzing and fusing the information from face and body pose. The method proposed in [17] used a I3D [7] backbone to extract spatio-temporal features followed by RNNs as the classifier.

Rather than using the standard combination framework like CNN-RNN architectures, in this work, we propose to leverage the recent powerful graph convolutional networks (GCNs) for seizure classification. The GCN model [20], which operates convolution on graphs, have been adopted in

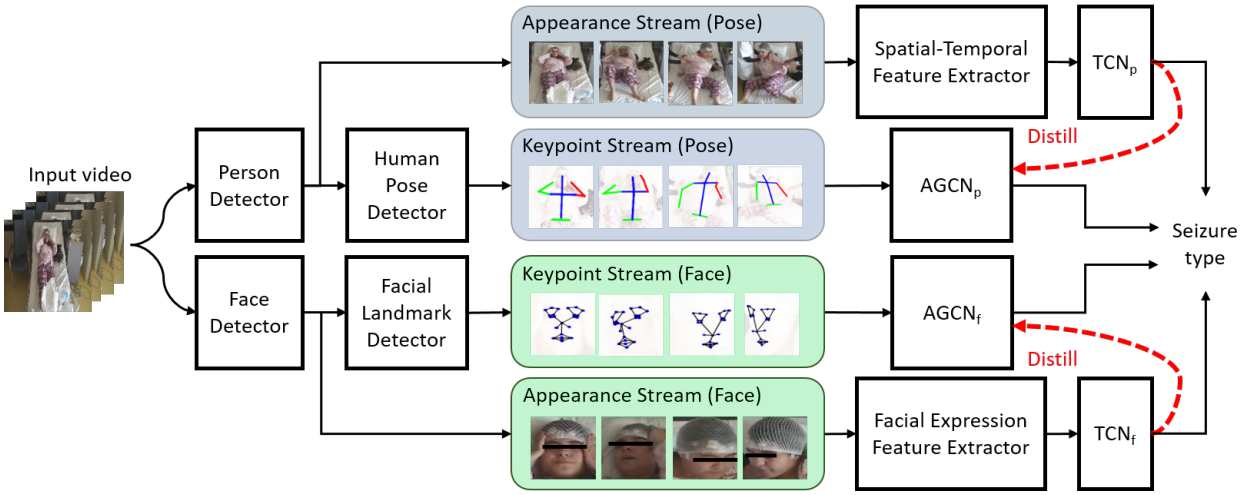

Figure 1. Overview of the proposed framework. In the training phase, knowledge distillation is applied from a trained TCN to regulate the learning of the corresponding AGCN. In the testing phase, only the learned AGCNs are used for final prediction, as shown in Fig. 5

various tasks, such as skeleton-based human action recognition [32, 29, 39] and facial landmark-based emotion recognition [27, 37, 38]. In this study, we apply an adaptive GCNs (AGCN) [32], in which the topology of the graph can be learned, on the detected body joints and facial landmarks for seizure classification. In addition, inspired by [29], we introduce a knowledge distillation (KD) mechanism from the complementary appearance stream for regulating the keypoint features learned by AGCN. To obtain further improvement, we combined the prediction from each AGCN separately trained on body pose keypoints and facial landmarks with the knowledge distillation mechanism. To our best knowledge, this work is the first attempt to utilize GCNs for seizure type classification (ES versus PNES) based on semiological information. The next section will describe the proposed methodology, followed by experimentation and conclusion.

2. Methodology

2.1. Overview

In this section, we describe our proposed multi-stream framework for classifying two types of seizures, i.e. ES and PNES. The overall architecture is shown in Fig. 1. After converting the seizure video into an image sequence, we detected and cropped the region of patient's body and face, followed by keypoint detectors for joint and facial landmark localization. The detected keypoints were then fed into separated AGCN for classification, which are viewed as Keypoint Streams. The cropped detected region of patient and face were fed into their corresponding feature extractor, and adopted temporal convolutional networks (TCNs) for temporal reasoning. The outputs of these streams, termed as Appearance Streams, were then used to transfer the learned knowledge to the Keypoint Streams. The predictions by AGCN from the pose and face streams were further combined for better performance. The following are the details for each stream.

2.2. Region of interest and keypoint detection

We adopted a fast SSD network [23] with MobileNet [15] backbone for region of interest (ROI) detection, i.e. detecting patients and their faces. The SSD model was pretrained on Imagenet dataset [10] and fine-tuned on our dataset. For body joint localization, we detected the 2D keypoints of upper-limb on the detected patient with Keypoint-RCNN [13, 25], which is pretrained on MS COCO [22] and fine-tuned on our dataset. The 11 detected points include head, neck, left/right shoulders, left/right elbows, left/right wrists, left/right hips, and bottom of the spine. The detected 2D keypoints were fed into a 3D estimator [9] for 3D pose estimation. For face stream, we used a toolbox [6] for extracting 2D facial landmarks with the detected face. There are 23 keypoints detected for each face, focusing on eyebrows, eyes, nose, and mouth. The toolbox was not optimized for our dataset. Fig. 2 and Fig. 3 show some illustrations and detection results.

2.3. The appearance stream

After the ROI detection on a video with T frames, we have the detected cropped region for patient as $R_P = \{r_{p1}, r_{p2}, \cdots, r_{pT}\}$ and for detected face as $R_F = \{r_{f1}, r_{f2}, \cdots, r_{fT}\}$, with $r_{pt} \in \mathbb{R}^{W_p \times H_p \times 3}$ and $r_{ft} \in \mathbb{R}^{W_f \times H_f \times 3}$. W_p and W_f are normalized width, and H_p and H_f are normalized height for detected regions for pose and face streams respectively. We leverage pretrained models for feature extraction followed by a temporal convolution layer. For pose stream, we used R(2+1)D model [35]

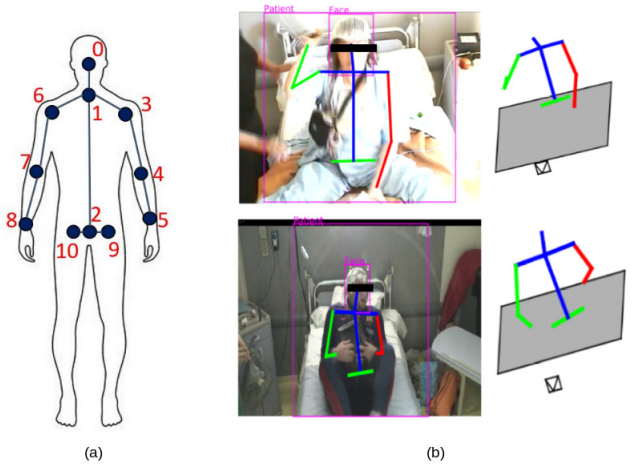

Figure 2. (a) Illustration of detected upper-limb joints. (b) Samples of ROI detection and (2D/3D) upper-limb keypoints detection.

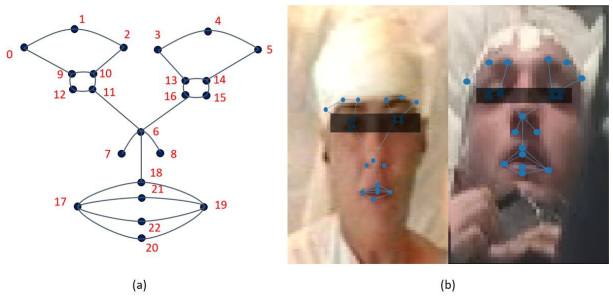

Figure 3. (a) Illustration of detected facial landmarks. (b) Samples of facial keypoint detection on our dataset.

pretrained on Kinetics [8] with the last classification layer removed as backbone to extract spatio-temporal features on a L-frame snippet, by

$$v_t = Model_{R(2+1)D}(r_{pt}, r_{p(t+1)}, \cdots, r_{p(t+L-1)}) \quad (1)$$

Hence for each time step t, the feature represents spatio-temporal information from a video snippet rather than a still image. As for facial feature extraction, we use the last layer output before classification layer of a VGG-19 model [33] pretrained on a public facial expression recognition dataset [1] as

$$u_t = VGG(r_{ft}) \quad (2)$$

With the extracted spatio-temporal feature sequence $V = \{v_1, v_2, \cdots, v_T\}$ and facial feature sequence $U = \{u_1, u_2, \cdots, u_T\}$, we feed them into respective temporal convolutional networks for video-level temporal reasoning as the following,

$$c_p = softmax(TCN_p(V)) \quad (3)$$

$$c_f = softmax(TCN_f(U)) \quad (4)$$

TCN_p and TCN_f represent the TCNs used for the pose and face streams respectively. Both of them are composed

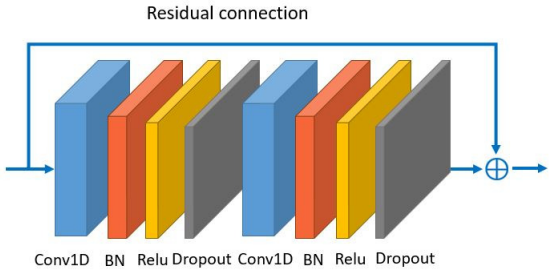

Figure 4. Illustration of the temporal convolutional block. Conv1D represents the 1D convolution on the temporal axis, followed by a batch normalization (BN) layer, a ReLU layer, and a Dropout layer. Moreover, a residual connection was added for each block.

by stacks of the temporal convolutional block, as shown in Fig. 4, followed by a linear layer as the classification layer. TCN_p and TCN_f are trained separately with standard cross-entropy loss for seizure classification. Later we used these pretrained models as teacher models to distill the learnt knowledge to the Keypoint Branches.

2.4. The keypoint stream

In the keypoint streams, we processed the spatio-temporal dynamics of detected keypoints for pose and face with their respective AGCN. The used AGCN is the one proposed in [32], in which the topology of the graph can be optimized while training for specific tasks. This property hence increases the flexibility of the model for graph construction and brings more generality to adapt to various data samples, such as the highly complex behavioral patterns in our case. For pose stream, we have detected upper-limb keypoint sequence $K_P = \{k_{p1}, k_{p2}, \cdots, k_{pT}\}$, with $k_{pt} \in \mathbb{R}^{C_p \times V_p}$ where C_p and V_p represent the number of channels and joints respectively. With pre-defined spatial adjacency matrix $A_p \in \mathbb{R}^{V_p \times V_p}$, describing the connection relation between the keypoints, we have output logits after $softmax$ operation as

$$o_p = AGCN_p(K_P, A_p) \quad (5)$$

Likewise for face stream, we have a facial landmark sequence $K_F = \{k_{f1}, k_{f2}, \cdots, k_{fT}\}$, with $k_{ft} \in \mathbb{R}^{C_f \times V_f}$ where C_f and V_f represent the number of channels and facial landmarks respectively. With the spatial adjacency matrix $A_f \in \mathbb{R}^{V_f \times V_f}$, we can have its output likewise by,

$$o_f = AGCN_f(K_f, A_f) \quad (6)$$

For the temporal dimension, we follow the paper [32], where each vertex is fixed as 2 (corresponding joints in the two consecutive frames). Instead of computing the cross-entropy for o_p and o_f, we introduced the learned knowledge in the Appearance Streams as addressed in the following part.

2.5. Knowledge distillation and ensemble

We have demonstrated how to process the appearance and keypoint information for both pose and face streams. For many multi-stream video analysis cases, it is usual to explicitly combine the learned knowledge from appearance and keypoint sources for a performance boost. Nevertheless, in this work we argue the keypoints should be the main information source for distinguishing seizures. First, we have decent fidelity of the keypoint detection throughout the whole videos. For the appearance stream, on the other hand, the occlusion often occurs in our dataset and so make the information less reliable. Besides, in medical scenarios like our study cases, privacy and confidentiality are important issues. To align these concepts, the strategy we adopted was to utilize both the appearance and keypoint information while training and only use keypoint information during testing. In addition to the cross-entropy loss, we introduced a standard knowledge distillation mechanism (KD) [14] while training the keypoint streams. It was implemented by minimizing the KL divergence between the probability distributions from the pretrained appearance streams and the keypoint streams. The overall objective losses for pose and face keypoint branches are hence as follows:

$$L_{CE,pose} = -\frac{1}{N}\sum_{i=1}^{N} y^i \cdot log(AGCN_p(K_p^i, A_p)) + (1-y^i)\cdot log(1-AGCN_p(K_p^i, A_p)) \tag{7}$$

$$L_{KD,pose} = \frac{1}{N}\sum_{i=1}^{N} D_{KL}(TCN_p(V^i)||AGCN_p(K_p^i, A_p)) \tag{8}$$

$$L_{CE,face} = -\frac{1}{N}\sum_{i=1}^{N} y^i \cdot log(AGCN_f(K_f^i, A_f)) + (1-y^i)\cdot log(1-AGCN_f(K_f^i, A_f)) \tag{9}$$

$$L_{KD,face} = \frac{1}{N}\sum_{i=1}^{N} D_{KL}(TCN_f(V^i)||AGCN_f(K_f^i, A_f)) \tag{10}$$

$$L_{Total,pose} = L_{CE,pose} + \lambda_p L_{KD,pose} \tag{11}$$

$$L_{Total,face} = L_{CE,face} + \lambda_f L_{KD,face} \tag{12}$$

,where $D_{KL}(P||Q) = \sum_j P_j log\frac{P_j}{Q_j}$, denoting the KL divergence. The λ_p and λ_f are trade-off hyper-parameters, and y^i is the label for the i-th example. We train the $AGCN_p$ and $AGCN_f$ separately. For the final prediction, we combined the prediction from pose and face streams for performance improvement, as shown in Fig. 5.

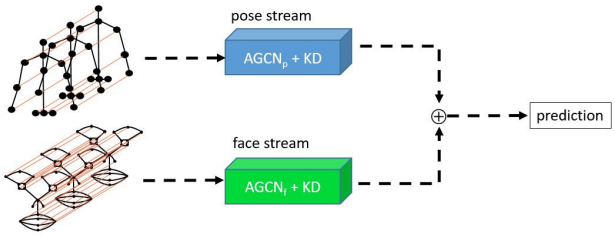

Figure 5. Illustration of the ensemble of the prediction from the pose and face streams in the testing phase, with the respective spatio-temporal graphs. The orange line denotes the temporal edges. AGCN+KD denotes AGCN network trained with additional knowledge distillation loss with the appearance streams as teachers.

Figure 6. Seizure examples in a real-world setting during daytime and night.

3. Experimentation

3.1. Dataset

In this work, we aimed to differentiate between ES and PNES, and tackle the problem in a real-world setting, as in Fig. 6, rather than a highly controlled environment. We collected 38 ES videos from 19 patients and 23 PNES videos from 15 patients, resulting in total 61 seizures and 34 patients. All patients have been recorded in the Video-EEG Epilepsy unit of the Epileptology department of the Timone University Hospital in Marseille, France. Both ES and PNES were selected according to presence of hyperkinetic motor behavior [5], which involve large amplitude, often explosive whole body movements. Due to the clinical challenges of localizing hyperkinetic ES seizures, and the challenges of discriminating between ES and PNES, this type of semiology is of great interest to neurologists [12, 19, 36]. The duration of the seizures ranged from 15 seconds to 180 seconds. Each patient had at least one and at most 6 recorded seizures. Both day and night conditions were included. All the seizure videos were collected from the video-EEG monitoring unit in the hospital. All patients had a firm diagnosis of either ES or PNES, established by expert epileptologists based on their video-EEG data. Patients gave informed consent for use of video-EEG data. Examples in Fig. 6 are from this dataset.

3.2. Data preprocessing

All seizure videos were converted to image sequence by 25 fps, and for each video, T frames were equally sampled for analysis. For video frame length shorter than T, the video itself was concatenated to enough frame length for sampling. In this study, T is set to 300. For image preprocessing, pixel values were normalized to 0 to 1.0, and normalized image size W_p, H_p, W_f, H_f are 112, 112, 48, and 48, respectively. For the 2D spatial coordinates of the detected keypoints, the values of the coordinates were normalized between -1.0 to 1.0 w.r.t the width and height of the cropped region. As for the third dimension in 3D pose estimation, the values were normalized with regards to the maximum and minimum values at the third axis across the video.

3.3. Quality of ROI and keypoint detection

As mentioned in section 2.2, we fine-tuned the ROI and keypoint detection with manually labeled data in our dataset. For ROI detection, the intersection-over-union (IoU) is used to for quantitative evaluation. The detection model used reached an average IoU of 0.89 for face detection and 0.94 for patient detection. As for the 2D body joint detection, the keypoint evaluation metric for MS COCO dataset is used. The mean average precision (mAP) at IoU of 0.50 is 0.67. As for facial landmark detection, the model used was not fine-tuned and we visually checked the quality of the results.

3.4. Experimental setup

We conducted both seizure-wise 10-fold cross validation and leave-one-subject-out validation on our datasets. Stochastic gradient descent (SGD) was applied as the learning optimizer. The initial learning rate for either of the four streams was 0.001, with linear learning rate decay scheduling used. The training epochs were set at 50, and we choose the weights at the epoch where the test sets had the highest accuracy for evaluation. The batch size was 4. The hyperparameters λ_p and λ_f are both set as 0.5, and the video snippet length L is 32. The configuration of $AGCN_p$ and $AGCN_f$ were the same as [32]. The kernel size and the dropout rate for both TCN_p and TCN_f are 4 and 0.4. The number of temporal convolutional blocks of the TCN for both pose and face streams are 6.

3.5. Experimental results

Table 1 shows the F1-score and accuracy of the 10-fold cross validation experiment, where

$$F1\text{–}score = \frac{TP}{TP + 0.5(FP + FN)} \tag{13}$$

model	F1-score	accuracy
$AGCN_p$	0.79	0.74
$AGCN_f$	0.78	0.70
TCN_p	0.75	0.69
TCN_f	0.80	0.74
$AGCN_p + KD$	0.86	0.84
$AGCN_f + KD$	0.84	0.82
Ensemble	**0.89**	**0.87**

Table 1. The 10-fold cross validation result: comparison of F1-score and accuracy between different models. AGCN+KD denotes AGCN network trained with additional knowledge distillation loss with the appearance streams as teachers.

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{14}$$

and TP, TN, FP, and FN denote true positive, true negative, false postive, and false negative, respectively. We take ES as a positive case. As shown in Table 1, we can see that $AGCN_P$ performs better than TCN_p, indicating that keypoint-based feature is more informative than appearance when correlating body pose to seizure classification. On the other hand, TCN_f slightly outperforms $AGCN_f$, inferring that for seizure analysis based on face, the appearance could provide more characteristic information than facial landmarks. Besides, for both the pose and face streams, we can have significant performance gain by introducing the knowledge distillation on the keypoint branch. This indicates the importance of utilizing complementary information (i.e. from keypoints and appearance) for seizure analysis. Lastly, combining the prediction from pose and face stream with our proposed ensemble method, the performances of the F1-score and the accuracy are 0.89 and 0.87, respectively. This performance improvement shows the effectiveness of integrating multi-stream information. Fig. 7 is the receiver operating characteristic (ROC) curve for different models in the 10-fold validation experiment. The ensemble model has the highest value of area under the ROC curve (AUC), indicating the best performance among the models. After the inclusion of knowledge distillation, AUCs of the keypoint branches can gain a significant boost. Table 2 shows the F1-score and accuracy of the leave-one-subject-out validation experiment. We can observe a performance drop compared to the 10-fold validation experiment, possibly due to that the inter-subject variance is considered in the setting and makes the task harder. Otherwise the overall result in Table 2 basically indicates the same trend and conclusion as that in the 10-fold cross validation. Besides, we also compare some deep learning based seizure classification studies with ours, as shown in Table 3 and Table 4. Table 3 shows how the methods in the related works performed in our task. Table 4 presents the results on their own work. Due to the limited number of studies using deep learning for video based seizure analysis, different seizure

model	F1-score	accuracy
$AGCN_p$	0.68	0.62
$AGCN_f$	0.68	0.59
TCN_p	0.53	0.56
TCN_f	0.68	0.61
$AGCN_p + KD$	0.74	0.67
$AGCN_f + KD$	0.72	0.66
Ensemble	**0.76**	**0.72**

Table 2. The leave-one-subject-out validation result: comparison of F1-score and accuracy between different models. AGCN+KD denotes AGCN network trained with additional knowledge distillation loss with the appearance streams as teachers.

model	F1-score (10-fold)	accuracy (10-fold)
[17]	0.80	0.71
[2](pose)	0.82	0.79
[2](face)	0.75	0.72
model	F1-score (LOSO)	accuracy (LOSO)
[17]	0.64	0.58
[2](pose)	0.70	0.62
[2](face)	0.66	0.61

Table 3. We implement the methods in Karácsony et al. [17] and Ahmedt-Aristizabal et al. [2], and test the model in our task. The table shows the results of 10-fold cross validation and leave-one-subject-out (LOSO) validation.

types are considered for comparison.

4. Conclusion

In this work, we propose a novel multi-stream framework with knowledge distillation for seizure classification, specifically for distinguishing between ES and PNES with hyperkinetic motor behavior. The contributions are twofold. First, we utilized multi-stream information from keypoint and appearance for both body pose and face streams. From experimental results, we give hints about which type of information should be used based on which stream information is being dealt with for seizure analysis, that is, for analysis based on body pose, keypoint-based features should be considered and for those based on face, appearance information seems more crucial. Second, by introducing a knowledge distillation mechanism, we show the importance of utilizing complementary information for keypoint-based seizure analysis. The performance obtained on real-world data for the challenging task of discriminating epileptic seizures from psychogenic non-epileptic seizures improve the state-of-the-art and are very encouraging with respective F1-score/accuracy 0.89/0.87 for seizure-wise cross validation and 0.75/0.72 for leave-one-subject-out validation.

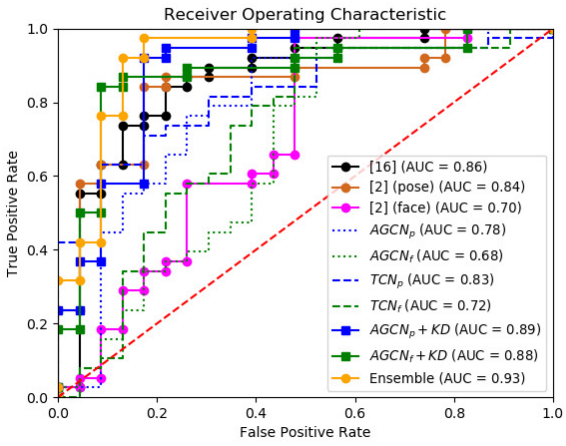

Figure 7. The 10-fold cross validation result: the ROC curve for the binary seizure classification task. AGCN+KD denotes AGCN network trained with additional knowledge distillation loss with the appearance streams as teachers.

Method	Classes	Performance
A.-Aristizaba et al. (2018) [2]	MTLE ETLE	Average accuracy: 0.53-0.56
Maia et al. (2019) [24]	TLE ETLE	AUC: 0.65
Karácsony et al. (2020) [17]	TLE FLE	F1-score: 0.84 AUC: 0.90
Ours	ES PNES	F1-score: 0.89 accuracy: 0.87 AUC: 0.93

Table 4. Comparison of deep learning-based seizure classification studies. The results shown are based on N-fold cross validation. MTLE, ETLE, and FLE denote mesial temporal lobe epilepsy, extra temporal lobe epilepsy, and frontal lobe epilepsy, respectively.

References

[1] Challenges in representation learning: Facial expression recognition challenge, 2013. 3

[2] D. Ahmedt-Aristizabal, C. Fookes, S. Denman, K. Nguyen, T. Fernando, S. Sridharan, and S. Dionisio. A hierarchical multimodal system for motion analysis in patients with epilepsy. *Epilepsy & Behavior*, 87:46–58, Oct. 2018. 1, 6

[3] D. Ahmedt-Aristizabal, C. Fookes, S. Dionisio, K. Nguyen, J. P. S. Cunha, and S. Sridharan. Automated analysis of seizure semiology and brain electrical activity in presurgery evaluation of epilepsy: A focused survey. *Epilepsia*, 58(11):1817–1831, Oct. 2017. 1

[4] D. Ahmedt-Aristizabal, K. Nguyen, S. Denman, S. Sridharan, S. Dionisio, and C. Fookes. Deep motion analysis for epileptic seizure classification. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, July 2018. 1

[5] W. T. Blume, H. O. Lüders, E. Mizrahi, C. Tassinari, W. V. E. Boas, and J. Engel. Glossary of descriptive terminology for ictal semiology: Report of the ILAE task force on classifi-

cation and terminology. *Epilepsia*, 42(9):1212–1218, Jan. 2002. 4

[6] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 2

[7] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. 1

[8] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017. 3

[9] C.-H. Chen and D. Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017. 2

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009. 2

[11] O. Devinsky, D. Gazzola, and W. C. LaFrance. Differentiating between nonepileptic and epileptic seizures. *Nature Reviews Neurology*, 7(4):210–220, Mar. 2011. 1

[12] J. Fayerstein, A. McGonigal, F. Pizzo, F. Bonini, S. Lagarde, A. Braquet, A. Trébuchon, R. Carron, D. Scavarda, S. Julia, I. Lambert, B. Giusiano, and F. Bartolomei. Quantitative analysis of hyperkinetic seizures and correlation with seizure onset zone. *Epilepsia*, 61(5):1019–1026, May 2020. 4

[13] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask r-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017. 2

[14] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. 4

[15] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. 2017. 2

[16] C. Hubsch, C. Baumann, C. Hingray, N. Gospodaru, J.-P. Vignal, H. Vespignani, and L. Maillard. Clinical classification of psychogenic non-epileptic seizures based on video-EEG analysis and automatic clustering. *Journal of Neurology, Neurosurgery & Psychiatry*, 82(9):955–960, May 2011. 1

[17] T. Karacsony, A. M. Loesch-Biffar, C. Vollmar, S. Noachtar, and J. P. S. Cunha. A deep learning architecture for epileptic seizure classification based on object and action recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2020. 1, 6

[18] Y. Kaya, M. Uyar, R. Tekin, and S. Yıldırım. 1d-local binary pattern based feature extraction for classification of epileptic EEG signals. *Applied Mathematics and Computation*, 243:209–219, Sept. 2014. 1

[19] A. Kheder, U. Thome, T. Aung, B. Krishnan, A. Alexopoulos, G. Wu, I. Wang, and P. Kotagal. Investigation of networks underlying hyperkinetic seizures utilizing ictal SPECT. *Neurology*, 95(6):e637–e642, July 2020. 4

[20] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017. 1

[21] W. C. LaFrance, G. A. Baker, R. Duncan, L. H. Goldstein, and M. Reuber. Minimum requirements for the diagnosis of psychogenic nonepileptic seizures: A staged approach. *Epilepsia*, 54(11):2005–2018, Sept. 2013. 1

[22] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft coco: Common objects in context. In *2014 ECCV*, 2014. 2

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. *Lecture Notes in Computer Science*, page 21–37, 2016. 2

[24] P. Maia, E. Hartl, C. Vollmar, S. Noachtar, and J. P. S. Cunha. Epileptic seizure classification using the NeuroMov database. In *2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG)*. IEEE, Feb. 2019. 6

[25] F. Massa. New person keypoint detection models in pytorch domain libraries, 2019. 2

[26] P. Maurel, A. McGonigal, R. Keriven, and P. Chauvel. 3d model fitting for facial expression analysis under uncontrolled imaging conditions. In *2008 19th International Conference on Pattern Recognition*. IEEE, Dec. 2008. 1

[27] Q. T. Ngoc, S. Lee, and B. C. Song. Facial landmark-based emotion recognition via directed graph neural network. *Electronics*, 9(5):764, May 2020. 2

[28] R. B. Pachori and S. Patidar. Epileptic seizure classification in EEG signals using second-order difference plot of intrinsic mode functions. *Computer Methods and Programs in Biomedicine*, 113(2):494–502, Feb. 2014. 1

[29] B. Pan, H. Cai, D. Huang, K. Lee, A. Gaidon, E. Adeli, and J. C. Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10867–10876. IEEE, 2020. 2

[30] K. Samiee, P. Kovacs, and M. Gabbouj. Epileptic seizure classification of EEG time-series using rational discrete short-time fourier transform. *IEEE Transactions on Biomedical Engineering*, 62(2):541–552, Feb. 2015. 1

[31] U. Seneviratne, D. Rajendran, M. Brusco, and T. G. Phan. How good are we at diagnosing seizures based on semiology? *Epilepsia*, 53(4):e63–e66, Jan. 2012. 1

[32] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019. 2, 3, 5

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 3

[34] A. Singh and S. Trevick. The epidemiology of global epilepsy. *Neurologic Clinics*, 34(4):837–847, Nov. 2016. 1

[35] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018. 2

[36] L. Vaugier, A. McGonigal, S. Lagarde, A. Trébuchon, W. Szurhaj, P. Derambure, and F. Bartolomei. Hyperkinetic motor seizures: a common semiology generated by two different cortical seizure origins. *Epileptic Disorders*, 19(3):362–366, Sept. 2017. 4

[37] C. Wu, L. Chai, J. Yang, and Y. Sheng. Facial expression recognition using convolutional neural network on graphs. In *2019 Chinese Control Conference (CCC)*. IEEE, July 2019. 2

[38] X. Xu, Z. Ruan, and L. Yang. Facial expression recognition based on graph neural network. In *2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC)*. IEEE, July 2020. 2

[39] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 2